Azzolini Riccardo 2020-04-30

Distribuzione della media di un campione

1 Calcoli di media e varianza su un campione

Sia X una variabile aleatoria che rappresenta i valori assunti da una qualche caratteristica quantitativa degli elementi di una popolazione di media μ e varianza σ^2. Estraendo da tale popolazione un campione di ampiezza (numero di elementi) n, si individuano le variabili $X_1, \ldots, X_n$, corrispondenti ai valori degli elementi estratti.

Se l'estrazione del campione è casuale, le $X_1, \ldots, X_n$ sono variabili aleatorie, i cui valori cambiano casualmente al variare del campione, e seguono tutte la stessa distribuzione di X (perché sia X che $X_1, \ldots, X_n$ rappresentano i valori di elementi casuali della popolazione).

1.1 Media campionaria

Dato un campione casuale (ancora di ampiezza n), la media dei valori dei suoi elementi è detta **media campionaria**:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Essendo calcolata a partire dalle variabili aleatorie X_i, la media campionaria $\overline{X}$ è essa stessa una variabile aleatoria (il cui valore cambia al variare del campione considerato), e dunque ha una propria distribuzione, che può essere studiata.

Allora, ha senso calcolare il valore medio $\mu_{\overline{X}} = E(\overline{X})$ della media campionaria su tutti i campioni: sfruttando la linearità del valore medio,

$$\begin{aligned} \mu_{\overline{X}} = E(\overline{X}) &= E\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(X_i) \end{aligned}$$

e il fatto che le X_i abbiano la stessa distribuzione di X,

$$\begin{aligned} &= \frac{1}{n} \sum_{i=1}^{n} E(X) \\ &= \frac{1}{n} \cdot nE(X) \\ &= E(X) = \mu \end{aligned}$$

si dimostra che il valore medio $\mu_{\overline{X}}$ della media campionaria è uguale alla media μ della popolazione (come, del resto, ci si aspetterebbe: i campioni vengono studiati proprio perché riproducono il comportamento dell'intera popolazione). Perciò, la media campionaria può essere usata come *stima* della media della popolazione (qualora quest'ultima non sia nota).

Analogamente, si può calcolare la varianza della media campionaria:

$$\begin{aligned} \sigma^2_{\overline{X}} = \operatorname{Var}(\overline{X}) &= \operatorname{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) \\ &= \frac{1}{n^2} \operatorname{Var}\left(\sum_{i=1}^{n} X_i \right) \end{aligned}$$

Se l'estrazione di ciascun elemento del campione non modifica la popolazione (perché è un'estrazione con reimmissione / rimpiazzo, oppure perché la popolazione è infinita), le X_i sono tra loro indipendenti, quindi vale la linearità della varianza:

$$\begin{aligned} &= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X) && (X_i \text{ e } X \text{ hanno la stessa distribuzione}) \\ &= \frac{1}{n^2} \cdot n \operatorname{Var}(X) \\ &= \frac{1}{n} \operatorname{Var}(X) = \frac{\sigma^2}{n} \end{aligned}$$

Quindi, mentre $\mu_{\overline{X}} = \mu$, la varianza $\sigma^2_{\overline{X}}$ della media campionaria *non* è uguale alla varianza σ^2 della popolazione, ma vale invece la relazione

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

purché il campionamento sia con reimmissione o la popolazione sia infinita.

Questa relazione implica che la varianza della media campionaria diminuisce all'aumentare della dimensione del campione. Infatti, intuitivamente, per campioni più grandi la media campionaria tenderà ad assumere valori più vicini al suo valore medio μ, cioè a fornire stime più accurate della media della popolazione.

1.2 Varianza campionaria a media nota

Se è nota la media μ della popolazione, si può calcolare la varianza dei dati di un campione rispetto a μ, detta **varianza campionaria a media nota** (da non confondere con la varianza della media campionaria):

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

Anche questa è una variabile aleatoria, che assume valori casuali cambiando campione, quindi si può calcolare il suo valore medio su tutti i campioni:

$$\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) &= \frac{1}{n}\sum_{i=1}^{n}E\big((X_i-\mu)^2\big) && \text{(linearità del valore medio)}\\
&= \frac{1}{n}\sum_{i=1}^{n}E\big((X-\mu)^2\big) && (X_i \text{ e } X \text{ hanno la stessa distribuzione})\\
&= \frac{1}{n}\cdot nE\big((X-\mu)^2\big)\\
&= E\big((X-\mu)^2\big)\\
&= \mathrm{Var}(X)=\sigma^2 && \text{(definizione di varianza)}
\end{aligned}$$

Quindi, la varianza campionaria a media nota fornisce una stima della varianza della popolazione, essendo in media uguale a quest'ultima.

1.3 Varianza campionaria a media incognita

Se, invece, la media della popolazione non è nota, si può calcolare la varianza dei valori di un campione rispetto alla media campionaria.

Se si applicasse direttamente la definizione di varianza, si otterrebbe

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{1}{n}\sum_{j=1}^{n}X_j\right)^2$$

che è una variabile aleatoria, e si può dimostrare che il suo valore medio (su tutti i campioni) è

$$E\left(\frac{1}{n}\sum_{i=1}n(X_i-\overline{X})^2\right)=\frac{n-1}{n}\sigma^2$$

Allora, la **varianza campionaria** (**a media incognita**) viene invece definita con il "fattore di correzione" $\frac{n}{n-1}$,

$$S^2=\frac{n}{n-1}\cdot\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

di modo che sia in media uguale alla varianza σ^2 della popolazione.

2 Esempio: simulazione numerica

Per osservare la distribuzione della media campionaria, si presenta in seguito una piccola simulazione numerica.

Si consideri una popolazione finita di $N = 4$ elementi, rappresentata da una variabile aleatoria discreta X avente la seguente distribuzione uniforme:

x_i	1	2	3	4
$f(x_i)$	0.25	0.25	0.25	0.25

La media μ e la varianza σ^2 di questa popolazione sono:

$$\begin{aligned}
\mu = E(X) &= \sum_{i=1}^{4} x_i f(x_i) \\
&= 1 \cdot 0.25 + 2 \cdot 0.25 + 3 \cdot 0.25 + 4 \cdot 0.25 \\
&= 0.25(1 + 2 + 3 + 4) \\
&= 0.25 \cdot 10 \\
&= 2.5
\end{aligned}$$

$$\begin{aligned}
\sigma^2 = \mathrm{Var}(X) &= E(X^2) - (E(X))^2 \\
&= 1^2 \cdot 0.25 + 2^2 \cdot 0.25 + 3^2 \cdot 0.25 + 4^2 \cdot 0.25 - 2.5^2 \\
&= 0.25(1 + 4 + 9 + 16) - 6.25 \\
&= 0.25 \cdot 30 - 6.25 \\
&= 7.5 - 6.25 \\
&= 1.25
\end{aligned}$$

Adesso, si vuole studiare la media campionaria per tutti i possibili campioni di ampiezza $n = 2$ estraibili da questa popolazione, con e senza reimmissione.

2.1 Campionamento con reimmissione

Se gli elementi dei campioni vengono estratti con reimmissione, i possibili campioni sono $N^n = 4^2 = 16$. Elencandoli tutti, è possibile calcolare il valore assunto dalla media campionaria

$$\overline{X} = \frac{1}{2} \sum_{i=1}^{2} X_i = \frac{X_1 + X_2}{2}$$

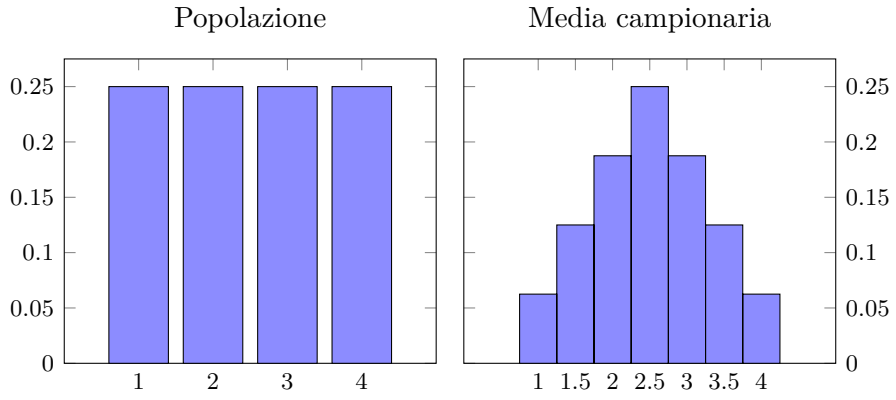
per ciascuno di essi:

Campione	Media	Campione	Media
$(1,1)$	$\frac{1}{2}(1+1)=1$	$(3,1)$	2
$(1,2)$	$\frac{1}{2}(1+2)=1.5$	$(3,2)$	2.5
$(1,3)$	2	$(3,3)$	3
$(1,4)$	2.5	$(3,4)$	3.5
$(2,1)$	1.5	$(4,1)$	2.5
$(2,2)$	2	$(4,2)$	3
$(2,3)$	2.5	$(4,3)$	3.5
$(2,4)$	3	$(4,4)$	4

Siccome l'estrazione di ciascuno dei campioni è equiprobabile (perché la popolazione è distribuita uniformemente), dalle frequenze (relative) dei valori assunti dalla media campionaria (il numero di occorrenze di ciascuno, diviso il numero totale di campioni, 16) si ricava direttamente la sua distribuzione:

$\bar{x}_i$	1	1.5	2	2.5	3	3.5	4
$f(\bar{x}_i)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Confrontando gli istogrammi delle distribuzioni della popolazione e della media campionaria,



si osserva che la distribuzione della media campionaria ha una forma simile alla campana della normale, anche se la popolazione è distribuita uniformemente.

Si passa poi al calcolo del valore medio della media campionaria,

$$\begin{aligned}\mu_{\overline{X}} = E(\overline{X}) &= \sum_{i=1}^{16} \overline{x}_i f(\overline{x}_i) \\ &= 1 \cdot \frac{1}{16} + 1.5 \cdot \frac{2}{16} + 2 \cdot \frac{3}{16} + 2.5 \cdot \frac{4}{16} + 3 \cdot \frac{3}{16} + 3.5 \cdot \frac{2}{16} + 4 \cdot \frac{1}{16} \\ &= \frac{1}{16}(1 + 3 + 6 + 10 + 9 + 7 + 4) \\ &= \frac{1}{16} \cdot 40 \\ &= 2.5 = \mu\end{aligned}$$

che, come previsto, coincide con la media della popolazione. Allo stesso modo, calcolando la varianza di $\overline{X}$,

$$\begin{aligned}\sigma^2_{\overline{X}} = \mathrm{Var}(\overline{X}) &= E(\overline{X}^2) - (E(\overline{X}))^2 \\ &= \frac{1}{16}(1^2 \cdot 1 + 1.5^2 \cdot 2 + 2^2 \cdot 3 + 2.5^2 \cdot 4 + 3^2 \cdot 3 + 3.5^2 \cdot 2 + 4^2 \cdot 1) - 2.5^2 \\ &= \frac{1}{16}(1 + 4.5 + 12 + 25 + 27 + 24.5 + 16) - 6.25 \\ &= \frac{1}{16} \cdot 110 - 6.25 \\ &= 6.875 - 6.25 \\ &= 0.625\end{aligned}$$

si vede che vale effettivamente la relazione

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{1.25}{2} = 0.625$$

2.2 Campionamento senza reimmissione

In alternativa, si può effettuare un campionamento senza reimmissione. Allora, i possibili campioni sono tutti i sottoinsiemi di $n = 2$ elementi della popolazione, che sono in totale

$$\binom{N}{n} = \binom{4}{2} = \frac{4!}{(4-2)!\,2!} = \frac{4 \cdot 3}{2} = 6$$

Come prima, elencandoli e calcolandone le medie,

Campione	Media
$\{1, 2\}$	1.5
$\{1, 3\}$	2
$\{1, 4\}$	2.5
$\{2, 3\}$	2.5
$\{2, 4\}$	3
$\{3, 4\}$	3.5

si ricava la distribuzione della media campionaria $\overline{X}$:

$\overline{x}_i$	1.5	2	2.5	3	3.5
$f(\overline{x}_i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

In questo caso, il valore medio della media campionaria è ancora uguale alla media della popolazione:

$$\begin{aligned}\mu_{\overline{X}} = E(\overline{X}) &= \frac{1}{6}(1.5 + 2 + 2.5 \cdot 2 + 3 + 3.5) \\ &= \frac{1}{6} \cdot 15 = 2.5 = \mu\end{aligned}$$

Invece, per la varianza della media campionaria,

$$\begin{aligned}\sigma^2_{\overline{X}} = \mathrm{Var}(\overline{X}) &= E(\overline{X}^2) - (E(X))^2 \\ &= \frac{1}{6}(1.5^2 + 2^2 + 2.5^2 \cdot 2 + 3^2 + 3.5^2) - 2.5^2 \\ &= \frac{1}{6}(2.25 + 4 + 12.5 + 9 + 12.25) - 6.25 \\ &= \frac{1}{6} \cdot 40 - 6.25 \\ &= \frac{20}{3} - \frac{25}{4} \\ &= \frac{80 - 75}{12} \\ &= \frac{5}{12}\end{aligned}$$

non si ha più la relazione

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

perché le variabili aleatorie X_i del campione *non sono indipendenti*, a causa dell'estrazione senza reimmissione. Vale però una relazione simile, ottenuta moltiplicando per un fattore di correzione $\frac{N-n}{N-1}$:

$$\begin{aligned}\sigma^2_{\overline{X}} &= \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \\ &= \frac{1.25}{2} \cdot \frac{4-2}{4-1} = \frac{1.25 \cdot 2}{2 \cdot 3} = \frac{1.25}{3} = \frac{5}{12}\end{aligned}$$

Osservazione: Se la popolazione è molto più grande del campione, il fattore correttivo $\frac{N-n}{N-1}$ è molto vicino a 1, quindi lo si può trascurare. Infatti, estraendo senza reimmissione pochi elementi da una popolazione grande, la composizione della popolazione non cambia significativamente, ovvero il campionamento può essere considerato approssimativamente con reimmissione. Al limite, su una popolazione infinita, il fattore correttivo vale 1, e non c'è alcuna distinzione tra campionamenti con o senza reimmissione.

3 Distribuzione della media campionaria

L'istogramma della media campionaria nell'esempio precedente suggerisce l'esistenza di un legame tra essa e la distribuzione normale.

Se si partisse da una popolazione distribuita normalmente, con media μ e varianza σ^2, la media campionaria $\overline{X}$ sarebbe a sua volta una variabile normale, in quanto distribuita attorno alla media della popolazione, con errori simmetrici. In particolare, $\overline{X}$ sarebbe normale di parametri

$$\mu_{\overline{X}} = \mu \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

Invece, nel caso generale, partendo da una popolazione non normale di dimensione N, media μ e varianza σ^2, la distribuzione della media campionaria ha valore medio $\mu_{\overline{X}} = \mu$, varianza

$$\sigma^2_{\overline{X}} = \begin{cases} \dfrac{\sigma^2}{n} & \text{se } \dfrac{n}{N} \leq 0.5 \text{ (regola pratica)} \\ & \text{o se il campionamento è con reimmissione} \\ \dfrac{\sigma^2}{n} \cdot \dfrac{N-n}{N-1} & \text{altrimenti} \end{cases}$$

e, per campioni di ampiezza n sufficientemente grande (come regola pratica, per $n \geq 30$), tale distribuzione è comunque approssimativamente normale.

3.1 Teorema del limite centrale

La possibilità di approssimare la distribuzione della media campionaria con la normale anche per popolazioni non normali (purché, come già detto, il campione sia abbastanza ampio) è giustificata dal *teorema del limite centrale*.

Teorema: Sia data una popolazione avente media μ e varianza σ^2. Da essa, si estraggono campioni casuali di ampiezza n; sia $\overline{X}$ la media campionaria riferita a tali campioni. Allora,

$$Z = \frac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$$

è una variabile aleatoria la cui distribuzione tende, per $n \to +\infty$, alla normale standardizzata.

Quindi, la distribuzione di $\overline{X}$ tende a una normale di media $\mu = \mu_{\overline{X}}$ e varianza

$$\left(\frac{\sigma}{\sqrt{n}}\right)^2 = \frac{\sigma^2}{n} = \sigma^2_{\overline{X}}$$

(supponendo che non sia necessario applicare il fattore di correzione per il campionamento senza reimmissione).

4 Problema sulla distribuzione della media campionaria

Problema: La variabile aleatoria continua X ha media $\mu = 5$ e varianza $\sigma^2 = 25$. Si estrae un campione di $n = 100$ elementi da questa popolazione; determinare la probabilità che la media del campione sia maggiore di 5.4.

Soluzione:

La media campionaria $\overline{X}$ ha valore medio $\mu_{\overline{X}} = \mu = 5$, e, supponendo che la popolazione sia infinita (o che il campionamento sia con reimmissione), varianza

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{25}{100} = 0.25$$

Essendo verificata la regola pratica $n = 100 \geq 30$, per il teorema del limite centrale si può affermare che $\overline{X}$ ha una distribuzione approssimativamente normale. Allora, standardizzando

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sqrt{\sigma^2_{\overline{X}}}} = \frac{\overline{X} - 5}{\sqrt{0.25}} = \frac{\overline{X} - 5}{0.5} = 2(\overline{X} - 5)$$

e consultando le tavole, si può calcolare la probabilità richiesta dal problema:

$$\begin{aligned} P\{\overline{X} > 5.4\} &= P\{Z > 2(5.4 - 5)\} \\ &= P\{Z > 0.8\} \\ &= 1 - P\{Z < 0.8\} \\ &\approx 1 - 0.7881 \\ &= 0.2119 \end{aligned}$$

5 Cenni sui tipi di campionamento

Il campionamento usato fin qui è quello casuale semplice (con e senza reimmissione). Esistono però anche dei metodi deterministici (ad esempio strati, grappoli, o suddivisioni sistematiche) che possono essere usati come parte del processo di selezione dei campioni. Quando si usano questi ultimi, bisogna o accertarsi che siano equivalenti all'estrazione casuale, oppure tenerne conto nei calcoli statistici.